作者主题模型及其改进的方法与应用研究综述

■ 徐涵[1,2] 刘小平[1,2]
[1] 中国科学院文献情报中心 北京 100190
[2] 中国科学院大学经济与管理学院图书情报与档案管理系 北京 100190

摘要：［目的/意义］作者主题模型作为近年来计算机领域关注度较高的新型概率模型，在文本挖掘与自然语言处理等方向已有广泛应用。分析国内外作者主题模型及其改进的思路与应用，更好地把握其研究现状，以期为计算机、图书情报等相关领域科研人员提供参考。［方法/过程］本文选取Web of Science核心数据库、DBLP及中国知网(CNKI)数据库作为文献来源，通过制定检索规则、去重及人工判读等操作提炼出关于作者主题模型及其改进方法的文献集，从模型应用过程的视角，结合文献分析法对现有研究进行总结归纳。［结果/结论］通过分析发现，现有相关研究已形成较为完整的分析流程，且模型的改进角度、适用领域也日益多样化。但性能优化、模型评价指标的规范完善以及在图书情报领域的进一步应用等方面仍有待深入探索。

关键词：作者主题模型 主题演化 社区发现 模型评估
分类号：G250 TP391
DOI：10.13266/j.issn.0252-3116.2019.07.016

1 引言

计算机技术的飞速发展、社交媒体作用下社区的快速连接，同时，数据类型的多样化、数据量的指数级增长，使得科研人员对大规模数据的信息挖掘已不再满足于用单一主题信息来代表数据集的特征，更多地是想挖掘主题与用户的关联关系，进而更全面地揭示数据集所包含的信息，发现其背后规律。基于此，科研人员于2004年提出了一种可以挖掘文档集中隐含作者与主题关系的统计模型，即作者主题(Author-Topic，AT)模型[1]。

AT模型是潜在狄利克雷分布模型(Latent Dirichlet Allocation，LDA)的扩展，继承了LDA主题模型将高维度的词集合映射到低维度的主题空间，进而实现数据降维。同时，将数据集中的作者信息作为扩展元数据融入到原模型中，使得数据集中的作者与主题建立起关联，为挖掘数据集中隐含的"作者-主题"语义信息提供了可能，有助于科研人员更好地将作者及其所讨论的内容对应起来。

随着模型应用范围的不断扩大，任务的多元化使得简单的AT模型难以满足日益增长的需求，如对社交媒体数据分析时需要考虑时间属性等。基于此，科研人员充分利用AT模型良好的可扩展性，对其尝试改进，并取得了一定突破，使其适用范围更为广泛。

笔者通过前期文献调研，发现现有综述文献大多是对主题模型为源头进行总结归纳，但随着主题模型的发展，使得由各种主题模型的改进模型所组成的模型集合十分宽泛，却少有针对于以"作者-主题"为脉络的一系列改进模型及应用的综述。同时，大部分综述主要是对模型原理及改进的时间顺序进行总结归纳，而从模型具体应用过程的视角进行综述则相对较少。因此，本文将立足于国内外有关AT模型的现有研究，采用文献分析及归纳总结法，深入分析国内外AT模型及其改进模型的方法及应用的研究现状，以期为相关科研人员提供参考，拓宽研究思路。

2 数据来源及概况

2.1 数据来源

为保证数据覆盖全面，本文在外文数据库的选取上除Web of Science核心数据库外，还选取了计算机领域较权威的数据库DBLP(Digital Bibliography & Library Project)。两者收录的文献虽有重叠，但仍存在互补。中文数据库则选取了中国知网(CNKI)。具体数据获取过程如下：①在CNKI数据库中，以"主题='作者'

作者简介：徐涵(ORCID:0000-0002-0896-1637)，硕士研究生；刘小平(ORCID:0000-0002-3342-8041)，副研究员，博士，硕士生导师，通讯作者，E-mail:liuxp@mail.las.ac.cn。
收稿日期：2018-07-02 修回日期：2018-11-27 本文起止页码：135-145 本文责任编辑：王传清

*（‘主题模型’+‘LDA 模型’）OR SU =‘作者主题模型’OR SU =‘Author Topic 模型’”进行专业检索，文献类型限定为期刊、会议及学位论文；在 Web of Science 核心合集中，利用“主题 =（“Author *” AND（“Topic Model *” OR “Latent Dirichlet Allocation”））”进行高级检索，文献类型为 Article、Proceedings Paper 及 Review；在 DBLP 中以“Author Topic Model”为主题词进行检索。以上检索均不限时间跨度，检索日期为 2018 年 5 月 25 日，分别检索出 174 篇中文文献，360 篇外文文献。②对上述 3 个数据集进行清洗，剔除无效文献、删除重叠文献。③人工判读，选取符合主题的文献，最终筛选出 139 篇文献（96 篇外文文献，43 篇中文文献）作为本文重点分析对象。

2.2 数据概况

关于 AT 模型及其改进模型的 139 篇研究论文中，包含 3 篇综述类文献，46 篇方法研究类文献，83 篇应用类文献，7 篇在实验环节将 AT 模型作为对照组的文献。图 1 为 139 篇文献的年代分布。由图 1 可知，AT 模型自 2004 年被提出后，国内对该模型进行研究且有科研成果产出始于 2008 年[2-3]。从发文趋势上看，其被关注度整体呈上升趋势，相关研究在 2013 – 2017 年期间的关注度先后迎来两次小高峰。根据上述数据可以推测基于 AT 模型的相关研究在今后一段时间内仍将保持较高的热度。

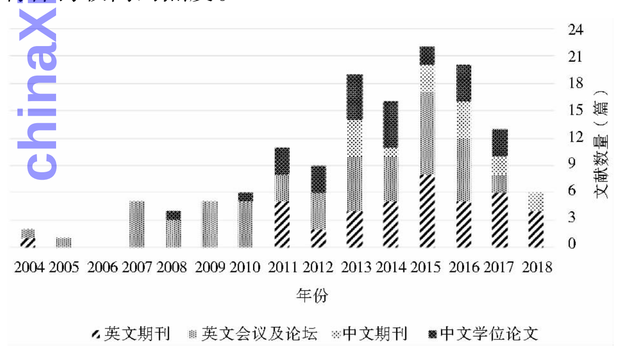


图 1 文献年代分布情况

3 研究进展分析

AT 模型的改进与应用研究的分析流程[4]如下（见图 2）：首先，确定待分析的原始数据集及数据类型；然后，对文本预处理结果进行建模并求解；最后，选择评价指标，通过对比实验对模型进行评估。本文将基于该流程对现有研究方法进行总结分析。

3.1 数据集选取

经过总结归纳，国内外基于 AT 模型的改进与应用研究所选用的数据集主要分为 6 类（见表 1）。其中，“学

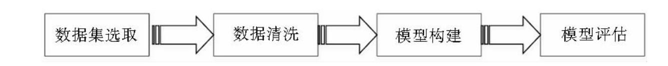


图 2 AT 模型及其改进方法与应用研究的分析流程

术论文”数据集类型数量位居首位（共 17 类），对其又可进一步细化为 5 小类：①专业领域数据库（DBLP、PubChem、MEDLINE、PubMed Central、ACM Digital Library）；②引文索引系统（Scopus、CiNii、NIPS、CiteULike、Citeseer、万方、CNKI、Web of Science）；③预印本文献库（Arxiv）；④学术搜索引擎（Arnetminer、Microsoft Academic Research）；⑤评测会议及语料库（TREC）。另外，文本类数据集中除学术论文这一大类外，还包括社交网站、邮件，而影像类数据集中除部分开放获取的数据外，还包含学校监控视频等非开放获取资源。

表 1 数据集类型分布

数据类型	子类	数据集名称	
文本类数据集	学术论文	DBLP[5]	TREC
		PubChem[6-9]	MEDLINE[10-11]
		Scopus[12]	PubMed Central
		CiNii[13]	万方[14]
		NIPS[15]	Arxiv[16]
		CiteULike	Web of Science[17]
		Citeseer[18]	CNKI[19]
		Arnetminer	ACM Digital Library
		Microsoft Academic Research[20-21]	
	社交网站	perverted - justice. com[22]	delicious. com
		IRC logs	ProgrammableWeb. com
		天涯论坛[23]	mooc. guokr. com[24]
		NLPIR	Yahoo
		新浪微博	Wikipedia[25]
		Twitter[26]	Tripadvis[27]
		digg[28]	
	邮件	Enron[29]	--
影像类数据集	开放获取	Flickr[30]	KTH
		VIPeR[31]	CASIA
		BrainMap	Weizmann
		DECODA[32-33]	IMDB
		Quickbird	Tripadvisor
		LIVE IQA	Google Earth[34]
	非开放获取	医院数据库	学校监控视频[35]

注：本分类是基于数据集在作者主题模型相关实验研究中，将被抽取的数据对象的类型进行归纳，而非数据集中所有的数据类型，如 Flickr 本身为社交平台，同时包含图像、文本类型的数据，但在 AT 模型所涉及到的实验中通常选取其图像数据进行分析，故被归为影像大类，而非文本大类

徐涵，刘小平. 作者主题模型及其改进的方法与应用研究综述[J]. 图书情报工作,2019,63(7):135－145.

确定数据集后,需要进一步提取目标数据字段。学术论文类数据集通常包括文献标题、作者、摘要及时间。针对不同研究内容,科研人员所提取的数据字段也存在一定差异。除上述 4 种字段外,部分研究还提取引文、期刊及会议信息等字段,以研究作者研究兴趣的变迁[36]、学术论文推荐[37]、学术领域专家社区发现[38]、特定学科领域内作者主题识别[39]等;对于影像类数据集,选取特定分辨率的图像及视频片段,实现图像及视频的主题识别[40－42],挖掘脑认知功能的空间对应关系[43]等;对于社交网站类数据集,通常提取推文正文,挖掘用户兴趣[44－49];对于邮件类数据集,通常提取收件人、邮件主题、内容、时间等字段,实现发件人与主题、时间的关联[50－52]。

3.2 数据清洗

获取目标数据字段后,需对其进行数据预处理。其中,作者消歧(author disambiguation)将对实验结果产生较大影响。而作者消歧目前面临两大难题:一是同名问题,即多人姓名相同;二是共指问题,即同一作者姓名的多种写法,该类问题常出现在英文文献中,与姓名缩写问题相关联。在实际操作中,大多是结合作者邮箱、工作单位及合著关系等属性特征来对作者进行消歧[53],同时可进一步借助外部网络信息提高消歧准确度[54]。具体可将作者消歧方法分为无监督学习与有监督学习两大类。其中,无监督学习方法包括图论与聚类,前者通过构建网络图将作者间的关系进行连接,计算节点间的拓扑距离,进而判断同名作者是否为同一个人。后者通过衡量相似度,将所有可能指向同一作者的同名作者聚为一类,而聚类方法的选择及同名作者间相似度函数的定义是影响聚类效果的关键[55]。有监督学习方法主要指基于概率模型的方法。该类方法通常需建立诸如贝叶斯网络、条件随机场等复杂概率模型,通过统计计算,推理得到重名作者间的匹配关系。

3.3 模型构建

完成数据清洗等操作后,根据研究内容及目标选择或构建合适的模型,并对模型进行参数估计。本节将对目前国内外 AT 模型及其改进模型进行总结与分析。

3.3.1 简单的 AT 模型　M. Steyvers 等于 2004 年提出了从文档集中挖掘作者与主题关联的方法,即作者主题模型[1],通过加入作者元素,同时对主题与作者进行建模。该模型将分析对象的作者信息引入到 LDA 模型中,以实现词、主题、作者及文档间的语义主题关联。每个作者可以通过多项分布矩阵映射至隐含主题上,每个主题的描述可通过词语层上的多项分布实现,文档可通过主题空间上的作者混合分布完成建模[56]。AT 模型的生成过程如下[57]:

对于一个文档集 D 中的每篇文档 $d \in D$:

第一,选取 θ(作者主题概率分布) ~ $Dirchlet(\alpha)$;

第二,选取 β(主题词概率分布) ~ $Dirchlet(b)$;

第三,对于 d 中的每个词 w:

(a)在 d 的作者集 a_d 中采用均匀分布以指定一个作者 x;

(b)根据作者主题概率分布 $x \sim Multi(x,\theta)$,指定一个主题 z;

(c)根据主题词概率分布 $z \sim Multi(z,\beta)$,指定一个词 w。

由该生成过程可知,AT 模型中的每位作者对应一个主题上的分布,且所有作者共享一个主题集合[58－60]。

3.3.2 扩展的 AT 模型　虽然作者信息的加入使 LDA 模型由无结构化信息改为结构化信息,使得主题模型可同时分析用户兴趣分布与文档结构,但依然存在一些缺陷[61]。近年来,随着文本挖掘技术的发展,科研人员进行了一系列的变形与扩展。下文将从基于时间因素、有监督学习方法、元数据扩展以及面向特定任务 4 个方面对基于 AT 模型的改进进行总结。

(1)基于时间因素的改进。随着 AT 模型的深入研究,如何得到作者所属的主题随时间变化的情况已成为当前研究热点。基于时间因素的改进从数据时间的角度可以进一步细化为两类:一是将时间视为随机变量,进而完成连续时间的建模;二是将时间离散化为一系列时间戳,对离散化的时间点构建动态贝叶斯网络。

- 连续时间建模。2010 年唐杰等人在 AT 模型中加入时间这一连续变量,提出了随时间变化的作者主题模型(Author-Topic-Time model,ATT)模型[49]。该模型将 AT 模型与随时间变化的话题模型(Time over Time,TOT)相结合,文档中每个词的生成由主题与时间两个属性共同决定,更好地描述了作者主题在不同时间的分布情况。

ATT 模型为连续时间建模的典型代表,后续有许多有关时间因素的作者主题挖掘研究是基于此开展的[62－65]。通过考虑主题之间的依赖关系、用户兴趣数据的稀疏性等问题,相继提出了基于隐马尔可夫的随时间变化的作者主题模型(Hidden Markov Author－

Time - Topic, HMATT)[49]、融入会议信息的随时间变化的作者主题模型(Author Conference Topic Over Time, ACTOT)[37]等模型,丰富了用户兴趣矩阵,降低了模型困惑度。

尽管 ATT 及 ACTOT 等模型相比于 AT 模型,考虑了文本的时间信息,以表示作者主题在不同时间的分布强度,但上述模型仍存在以下两个问题:第一,模型在每个时间窗口内主题数是固定的,因此只能揭示主题强度的变化,而忽略了主题内容的变化[58];第二,对于一些随时间变化较为频繁、规模较大的文档集,模型消耗大量计算及内存资源。

- 离散化时间建模。针对 ATT 模型无法实现对动态数据的分析及主题内容追踪等不足,离散化时间建模的思想应运而生。以 2015 年杨如意提出的动态作者主题模型(Dynamic Author Topic Model, DAT)[66-67]、2018 年余传明等提出的复合主题演化模型(Author Topic Time-Latent Dirichlet Allocation with Author Ranking, ATT-LDA)[68]为代表,该类模型通常结合了动态主题模型(Dynamic Topic Model, DTM)及 AT 模型两者优点,将获取的主题离散到时间窗,使其在计算复杂度和应用场景两方面有较大优势。经实验证明该类模型能准确描述潜在作者主题及其动态变化。

(2)基于有监督学习方法的改进。目前大多基于 AT 模型扩展的模型均为无监督模型,通常只需要输入文档集、主题数目等数据参数,模型即可实现自动学习。然而,无监督学习方法得到的结果通常可解释性较差,甚至不易被理解。针对该问题,2015 年 H. Mou 等人提出了有监督作者主题模型(Author Subject Topic, AST)[14],通过引入有监督的"subject"层来对文档进行分组。与传统的 AT 模型相比,该类模型有助于将单词和文档聚在一起以减少噪声。

(3)基于元数据元素扩展的改进。引入更为丰富、多样化的元数据元素,将传统的"作者 - 主题"分布转变为"其他元数据 - 作者 - 主题"分布,进一步拓宽 AT 模型可揭示的内容范围。基于元数据扩展的改进目前主要分为以下 4 种:①融入社区元素,如 2012 年 C. S. Li 等提出的"作者 - 主题 - 社区" 模型(Author-Topic-Community, ATC)等[34,69-71],该类模型结合社会网络分析,根据作者兴趣实现作者社区发现;②融入会议元素,如 2008 年唐杰等提出的融入会议信息的模型(Author-Conference-Topic, ACT)及 2011 年 Y. Ding 提出的 ACTC(Author-Conference Topic-Connection)模型[3,72];③融入用户兴趣等用户特征元素[44,61,73],该类模型可以更好地应用于个性化推荐等任务中;④融入作者引文等相关元素,其中主要包括引文、所属期刊等[74-75],使得模型充分利用学术文档中作者和引文信息,对文档具有较高的作者判别和引文文献排序能力。

(4)面向特定任务的改进。为了让 AT 模型满足特定领域中的特定任务,科研人员提出了面向特定任务的 AT 模型。该类模型目前针对的数据大多是图像。其中,最具有代表性的是通过融入地理信息(Author - related Geographical Topic Modeling, AGTM)[6]及多光谱遥感图像的颜色、形状特征(Author - Genre Topic Model, AGTM)[41]对区域类别进行主题标注。但如何在结合地理信息的同时使用更优的概率模型来处理文档信息以提升图像类别标注的精准度仍需进一步研究。

上文所介绍的 4 类扩展模型主要是针对传统 AT 模型的现有问题,从不同角度改进而成的。图 3 对各模型的扩展模式进行总结归纳,主要分为以 LDA 及 AT 模型为基础的两大扩展模式,其中,实线箭头表示由箭尾一侧的模型改进成箭头所指一侧的模型,虚线箭头表示将虚线箭尾所指的模型与其作用的实线箭尾模型结合,进而得到实线箭头所指的模型。表 2 是对上文所介绍的 4 类扩展 AT 模型的汇总,同时对每个模型的优缺点等信息进行分析。

3.4 模型评估

衡量结果是否真实揭示数据规律以及相比于已有研究是否更为优化,在整个实验过程中必不可少。随着科学研究的不断推进,模型评价角度更为全面。经过调研发现,针对 AT 模型及其改进的研究中常用的模型评估方法主要从以下 6 个角度开展,具体模型评价指标分类见图 4。

3.4.1 *模型泛化能力* 通俗来讲是指训练后的模型对未知数据的预测能力,在实际操作中,通过测试误差来量化模型的泛化能力。目前衡量模型泛化能力最常用的度量指标是困惑度(perplexity)。在本文所构建的数据集中,Q. Q. Yang 及 L. Poddar 等利用对该指标对模型泛化能力进行量化评估,实验中困惑度表示主题词预测的难易程度,值越低说明泛化能力越好[78-79]。此外,万路康等利用 P@n(Precision@N)对模型泛化能力进行评估,用以检测前 N 条预测结果的准确率[80]。在实际应用中,常选取 P@5、P@10,及 P@20。

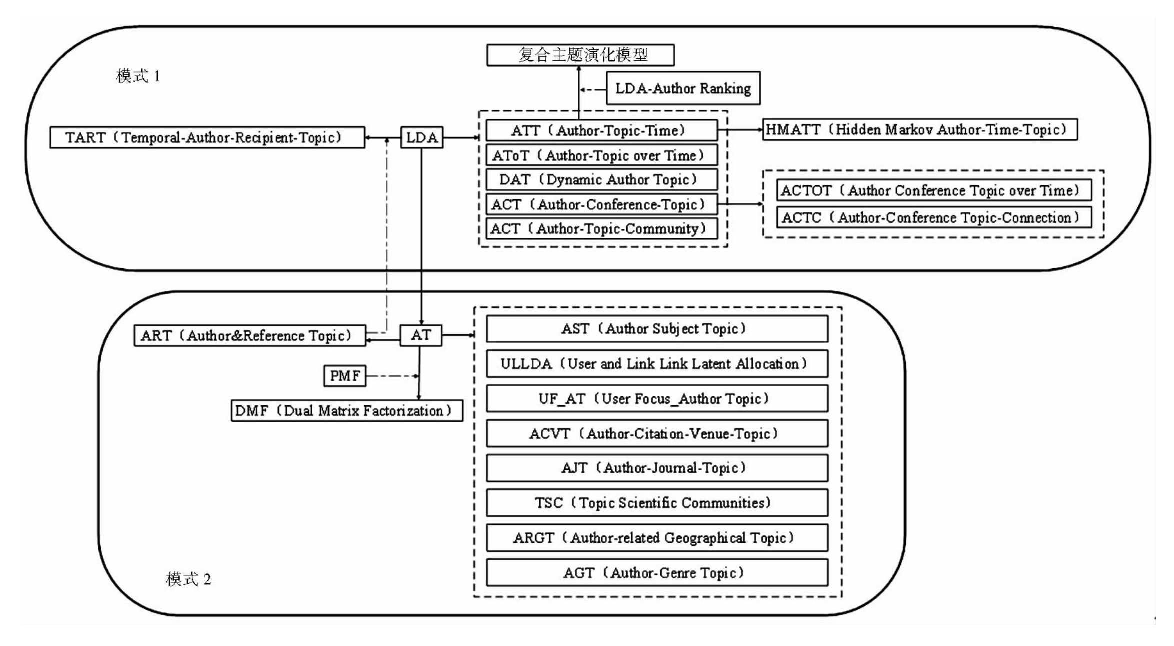


图 3　AT 模型及其改进模型的扩展模式

表 2　扩展的 AT 模型汇总

扩展类型	时间（年）	作者	模型	简单描述	优点	缺点
基于时间因素的改进	2005	A. Mccallum 等[76]	TART	结合作者、接收者和时间因素分析主题	发现用户间的关系	没有对时间段间的动态关系建模
	2010	唐杰等[49]	HMATT	可建立主题之间相互依存关系	实现作者－主题基于时间的演化分析	没有对时间段间的动态关系建模
	2011	N. Naveed 等[77]	ATT	结合作者和时间戳信息来捕捉用户兴趣随时间变化的潜在主题		
	2013	史庆伟等[36]	AToT	引入作者和时间元数据，描述主题演化规律及作者研究兴趣的变迁	考虑了文本的时间信息，可以表示主题在不同时刻的分布强度	每个时间窗内的主题数是固定的，只能揭示主题强度的变化趋势，忽略了主题内容的变化
	2014	刘智超等[37]	ACTOT	根据用户评分的论文内容、期刊会议及发表时间等信息对用户兴趣进行建模，以更加准确地描述用户兴趣特征	可以准确地对研究人员的兴趣进行建模	时间复杂度较高
	2015	杨如意等[66－67]	DAP	既考虑了时间因素作为内因对于主题变化的影响，又把握了数据集整体的主题内容	在应用场景和计算复杂度两个方面有所提升	文档主题服从多项分布，且不同主题间具备弱相关性，这与实际语料不是很符合
	2017	余传明等[68]	复合主题演化模型	结合 ATT 与 LDA-Author Ranking 两者优点	提升了作者－主题演化结果	未考虑权重对作者题强度变化的影响
基于有监督学习方法的改进	2015	H. Mou 等人[14]	AST	不同主题分布下的作者及文档共享主题信息	能够有效地捕捉主题类，有效地区分主题，为专家的研究兴趣建模，有助于专家推荐	时间复杂度较高
基于元数据元素扩展的改进	2008	J. Tang 等[3]	ACT	增加会议主题信息	加入会议信息，更为精准地提升了作者－主题的匹配程度	只能生成每个主题的作者概率分布，但同一主题下作者可能不同属一个社区，仍不能解决社区与主题之间的交互问题
	2012	J. Wang 等[38]	ACTC	增加会议主题及主题间潜在的映射信息	动态设置对应于不同会议的主题数量	设置主题数量的操作较复杂
		C. S. Li 等[69]	ATC	增加作者－主题、作者－社区的映射信息	模型结果较现有模型有所提升，如作者兴趣分析和作者社区发现	模型在一定程度上利用引文信息定义合作关系，使得社区构建中成员组成不够精准
	2013	J. Yang 等[73]	DMF	结合 PMF 与 ATM 算法对用户评分及主题词矩阵进行分解	可结合用户兴趣及文本潜在特点进行精准推荐	没有明确目标函数，且各通路权重靠迭代前后各通路消息差值累加来估计，缺乏理论证明

（续表 2）

扩展类型	时间（年）	作者	模型	简单描述	优点	缺点
		亓晓青等[61]	ULLDA	作者对应的主题分布不再只是由作者一人决定，而是根据微博的特点，使相关人员的主题分布也能影响该分布	结合微博数据特点，使结果更为精准	通过经验确定主题数目，缺乏考证
	2014	江雨燕等[74]	ART	以 USTM（Upstream Supervised Topic Model）和 DSTMDSTM（Downstream Supervised Topic Model）方式构建了文档作者和引用文献的生成过程	将引文信息及有监督模型融入其中	引用信息涉及一些无效引用等干扰
		Z. Yang 等[75]	ACVT	对作者、论文、被引作者等信息综合建模	包含更多有价值的上下文信息	
		M. Morchid 等[32]	TSC	通过分析 TSC 的网络结构，揭示主题上的协作模式	考虑了作者协作模式	参数设定依据有待验证
	2015	王永贵等[44]	UF_AT	微博特性与 AT 模型结合	主题词更加准确，关联性更强	影响微博用户兴趣的因素考虑不够全面
	2016	J. K. Ha 等[39]	AJT	兼顾作者与期刊主题进行分析	考虑了相邻引文句之间的上下文关系	传统的基于频率的引文分析方法适用性有待验证
面向特定任务的改进	2012	罗杰斯[30]	ARGT	引入地理区域的概念，得到主题 – 区域分布	适用于图像数据，拓宽了模型适用领域	在文档层面上没有提出好的概率模型
		W. Luo 等[41]	AGT	采用了一种适用于多光谱遥感图像的基于颜色和形状特征的特征描述符，结合 AT 模型，实现变化区域类别标注	提高了变化区域类别标注的精度	缺少适用的概率图模型

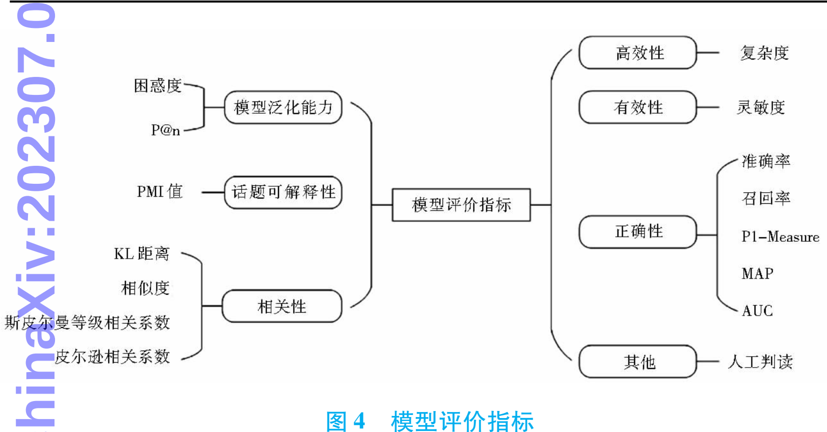


图 4　模型评价指标

3.4.2　话题可解释性　通过话题中单词分布的语义相关性进行分析，但实验证明，此评价角度具有较大主观性。为了相对客观地对其进行评价，有学者借助计算 K. M. Schneider 等提出的 PMI（pointwise mutual information）值进行量化，PMI 值常用以衡量一对事件共同出现的概率[81]。

3.4.3　高效性　目前大多数研究人员通过复杂度对模型的高效性进行评估[6]。复杂度具体指算法对应的程序运行时所需资源，可进一步分为时间复杂度和空间复杂度。但基于复杂度的评估方法无法避免的问题在于模型复杂度低并不能代表该模型在语义层面可以得到良好的主题词，因此，H. M. Wallach 等在已有评估方法的基础上，提出了包含有效性及正确性在内的两种不同评价维度[82]。

3.4.4　有效性　有效性是模型输出结构与真实生命系统的特性符合程度的定量表示。C. S. Li 等人曾利用灵敏度（sensitivity）刻画模型对单位待评价变量变化所致的响应量变化程度[69-70]。除此之外，W. Buntine 对 H. M. Wallach 的 left – to – right 算法进行改进，提出了一种新型估计文档集合的似然方法，使得模型评价更具有公平性[83]。

3.4.5　正确性　正确性包含查全率、查准率及相关综合评价指标。在本文构建的数据集中，王永贵等借助准确率（precision）、召回率（recall）及结合准确率 P 和召回率 R 的综合评价指标 F1 – Measure 对模型正确性进行定量分析[45]。此外，还有 X. Xie 等利用 AUC（Area Under the ROC Curve）指标衡量模型正确性，其中，ROC 曲线（纵轴：真正例率 TPR；横轴：假正例率 FPR）下的面积越大表示模型越精确[84]；J. Wang 等引用平均准确率这一概念（Mean Average Precison，MAP）评价模型正确性，其中，单个主题的平均准确率是每篇相关文档检索后准确率的平均值[38]。

3.4.6　相关性　该评价角度在实际操作中通常借助一些具体任务进而间接完成对模型的评估，具体而言是利用距离这一可量化的概念反映文档或作者间主题

徐涵，刘小平. 作者主题模型及其改进的方法与应用研究综述[J]. 图书情报工作,2019,63(7):135－145.

相似性等。如在本文所构建的数据集中，C. S. Li 等借助 KL 距离（Kullback－Leibler Divergence）[57]、T. Wang 等借助相似度（Similarity）[6]衡量相同空间内两个作者主题概率分布的差异情况；此外，还有 T. Zhang 等利用斯皮尔曼等级（Spearman Rank）相关系数及皮尔逊相关系数（Pearson Correlation Coefficients）对相关性进行评估[85]。其中，斯皮尔曼等级相关系数认为如果数据中没有重复值，且当两个变量完全单调相关时，值为＋1 或－1；而皮尔逊相关系数的取值范围通常在[－1，1]之间。

除上述 6 大类评价角度外，研究人员还提出一些弹性评价方法，如人工评价[46]。但这种评价方法的主观性较强，仍存在很多问题。

通过上述总结分析发现，不同评价角度对模型衡量的侧重点大不相同，而科研人员在对实验进行评估时往往会人为地根据程序所实现的任务选取相应评价指标，使得不同模型之间的横向比较变得困难。同时，现有研究表明，模型在特定评价指标下表现良好时，在其他方面的性能可能落差较大。以目前常见定量评估指标复杂度为例，若改进模型的复杂度低于现有模型，即表明改进模型的建模思想更为优化。但正如前文所提，该指标存在最大的问题是复杂度的高低并不能代表主题词挖掘结果的好坏[82]。因此，促成了有效性、正确性等众多评价指标的应运而生，尽可能使评价更为公平。但由于模型的应用领域、时间切片划分等差异，截至目前，模型之间的好坏判断依旧没有统一标准。

3.5 基于 AT 模型的应用

由于 AT 模型的研究不断深入，目前已被广泛应用到文本挖掘等众多领域。其中，受关注较多的领域包括社交媒体分析、学术文献以及社区发现等。下文将重点介绍 AT 模型及其改进方法在以上领域中的应用。

3.5.1 社交媒体中的应用　随着互联网的飞速发展，社交媒体也随之兴起。微博及 Twitter 作为新媒体的代表，相比传统文档数据而言，其数据更新速度快，且伴有网络用语，不够规范。因此，如何结合社交媒体数据的特点来完成数据分析任务是目前研究热点之一。

早期将 AT 模型应用至社交媒体数据分析主要是根据数据内容进行建模。但该思想忽略了社交媒体数据极强的实时性，未将文本内容与时态相结合，导致模型无法观察主题随时间的变化。此外，由于 Twitter 中文本内容长度过短，使得模型无法通过无监督学习获得较为理想的主题分布情况。C. S. Li 等提出的 ATC（Author-Topic-Community）模型解决了此问题，同时实现了作者兴趣及其社区结构的同步推断[57]。后续还有基于此进行用户兴趣挖掘[44,86]、特定领域微博账号萃取[87]、基于用户兴趣的微博推荐[88－89]等研究。

3.5.2 学术文献中的应用　在学术文献中，AT 模型最具有代表性的应用便是通过考虑时间因素，对不同时间段“作者－主题”对应分布予以揭示，深入理解前沿交叉学科带头人的主题演化过程及发展趋势，在一定程度上揭示主题间的相互影响，同时在预测研究趋势方面也有一定参考价值。此外，由于学术文献不同于博文，通常会附有参考文献、期刊会议等信息。因此，关于会议、引文等相关元素的融入也陆续引起科研人员的关注。如 Y. Tu 等于 2010 年提出了融入引用信息的模型（Citation-Author-Topic，CAT）[90]，该模型同时对论文作者及被引用的作者建模。后续，又有学者就作者影响力的问题进行了延伸思考[91]。

3.5.3 数据社区中的应用　目前，AT 模型在数据社区中的应用主要包括社区发现及推荐系统。社区发现根据分析数据类型的不同，可以划分为社交媒体及科研合作团队两种类型。利用社区发现方法可以有效地发现数据结构及演化过程，有助于科研人员分析数据集中存在的网络结构及其性质，进而了解整个网络的变化趋势，为资源搜索、推荐以及网络结构优化等提供服务[92]。同样，推荐系统也可分为社交媒体及科研领域两类数据，其推荐的依据均为用户/作者的兴趣。

3.5.4 其他　除上述主要应用领域外，还有一些根据科研人员具体实验内容所实现的特定应用领域。如图像领域的自动类别标注，该类实验通常选取特定分辨率的图像及视频片段，实现图像及视频的主题识别，较为典型的实验包括挖掘脑认知功能的空间对应关系等[43]；此外，还有对邮件数据进行分析，通过提取收件人、邮件主题、内容、时间等字段，实现发件人与主题、时间的关联等[50]。

4 研究不足与发展趋势

上文对 AT 模型的具体应用进行了总结。尽管 AT 模型已被不断改进，但由于需求及计算机技术的更新迭代，仍存在一些不足，具体体现在以下 4 个方面：

(1)技术层面:社交媒体的数据规模越来越庞大,并且其数据具有过短的文本长度、快速的更新速度等特点,进而在短时间内产生了大量网络用语及符号语言,使得数据的噪声增大、上下文信息获取不足。上述情况的产生会大大增加 AT 模型处理数据的难度,导致主题识别效果无法达到预期。

(2)特定领域内方法选择层面:AT 模型初衷是对科技文献类型的数据进行分析,实现"作者 - 主题"的关联,以揭示特定学科领域内的某些发展演化规律,且效果较为理想。但在后期应用中发现,AT 模型已在社交媒体乃至图像领域有着较为广泛的应用,而在图书情报领域的应用尚不成熟。大多研究更注重算法改进、算法性能提升,但图书情报领域中充分应用 AT 模型及其改进模型进行主题发现等相关研究的相对较少,特别是国内研究仍有部分停留在应用计量分析软件的层面。

(3)数据集选择层面:在 AT 模型应用于科技文献类型数据的研究中,大部分实验数据来自于语料库或公开数据集,侧重于验证模型的有效性。但考虑通过 JCR、中国科学院分区以及特定领域内权威论文、会议评价体系等方法,尽可能全面地确定领域具有代表性的数据集,揭示某一学科或领域演化的研究还不够。

(4)缺少模型优劣性评价体系:科研人员通常根据经验或任务特性选择相关评价指标,但所选指标能否客观地衡量模型优劣性依旧是 AT 模型领域所面临的主要问题之一,如何改进或提出新的评价指标将成为未来研究的重点。

综上所述,未来 AT 模型的相关研究可以向性能优化、模型评价指标体系的规范完善以及在图书情报领域的深度应用 3 个方面发展:性能优化体现在高效的训练算法,减少时间或空间复杂度等;模型评价体系具体可根据任务性质等角度进行划分,结合任务性质与评价指标的原理,尽可能在领域内对于不同任务的模型评价方法达成一致;在图书情报领域的进一步应用将是一大发展趋势,如挖掘用户兴趣特征,实现学者画像、图书馆或学者个性化与精准服务,或对前沿交叉学科的科研合作及主题演化进行研究以对管理决策部门起到支持作用等。

参考文献:

[1] STEYVERS M, SMYTH P, ROSEN-ZVI M, et al. Probabilistic author-topic models for information discovery[C]// The tenth ACM SIGKDD international conference on knowledge discovery and data mining. Seattle, Washington: ACM, 2004: 306 - 315.

[2] 骆国靖. 基于主题模型的模块化网络和社区挖掘研究[D]. 杭州:浙江大学, 2008.

[3] TANG J, ZHANG J, YAO L, et al. ArnetMiner: extraction and mining of academic social networks[C]// ACM SIGKDD international conference on knowledge discovery and data mining. Henderson: ACM, 2008: 990 - 998.

[4] 王燕鹏. 国内基于主题模型的科技文献主题发现及演化研究进展[J]. 图书情报工作, 2016, 60(3): 130 - 137.

[5] 吴良, 黄威靖, 陈薇, 等. ACT-LDA: 集成话题、社区和影响力分析的概率模型[J]. 计算机科学与探索, 2013, 7(8): 718 - 728.

[6] WANG T, HUANG Z, GAN C. On mining latent topics from healthcare chat logs[J]. Journal of biomedical informatics, 2016, 61(C): 247 - 259.

[7] MORCHID M, BOUAZIZ M, KHEDER W B, et al. Spoken language understanding in a latent topic-based subspace[C]// International symposium on computer architecture. San Francisco: Springer, 2016: 710 - 714.

[8] LEE M, HUANG R, TONG W. Discovery of transcriptional targets regulated by nuclear receptors using a probabilistic graphical model [J]. Toxicological sciences, 2016, 150(1): 64 - 73.

[9] XUAN J, LU J, ZHANG G, et al. Infinite author topic model based on mixed gamma-negative binomial process[C]// IEEE international conference on data mining. Atlantic City: IEEE, 2016: 489 - 498.

[10] 陈霄咚. 生物医学领域的专家寻找研究[D]. 上海:复旦大学, 2013.

[11] NEWMAN D, KARIMI S, CAVEDON L. Using topic models to interpret medline's medical subject headings[C]// Australasian joint conference on advances in artificial intelligence. Berlin, Heidelberg: Springer, 2009: 270 - 279.

[12] CHAIWANAROM P, ICHISE R, LURSINSAP C. Finding potential research collaborators in four degrees of separation[C]// Advanced data mining and applications, international conference, ADMA 2010. Chongqing: DBLP, 2010: 399 - 410.

[13] ICHISE R, FUJITA S, MURAKI T, et al. Research mining using the relationships among authors, topics and papers[C]// International conference information visualization. Zurich: IEEE Computer Society, 2007: 425 - 430.

[14] MOU H, GENG Q, JIN J, et al. An author subject topic model for expert recommendation [M]//Information retrieval technology. Cham: Springer International Publishing, 2015: 83 - 95.

[15] 薛维. 基于非对称先验的作者主题模型[D]. 杭州:浙江大学, 2011.

[16] KIM J, KIM D, OH A. Joint modeling of topics, citations, and topical authority in academic corpora[J]. Transactions of the Asso-

ciation for Computational Linguistics, 2017, 5(8):191 – 204.

[17] MAO J, CAO Y, LU K, et al. Topic scientific community in science: a combined perspective of scientific collaboration and topics[J]. Scientometrics, 2017, 112(2):851 – 875.

[18] 吴钟刚，吕钊. 一种基于局部相似性的社区发现算法[J]. 计算机工程, 2016, 42(12):196 – 203.

[19] 关鹏，王曰芬. 学科领域生命周期中作者研究兴趣演化分析[J]. 图书情报工作, 2016, 60(19):116 – 124.

[20] JEONG Y S, LEE S H, GWEON G. Discovery of research interests of authors over time using a topic model[C]// International conference on big data and smart computing. New York: IEEE Computer Society, 2016:24 – 31.

[21] FENG S, CAO J, CHEN Y, et al. A model for discovering unpopular research interests[C]// International conference on knowledge science, engineering and management. Cham: Springer, 2015: 382 – 393.

[22] KUZNETSOV A, KYPRIANOU A E, PARDO J C. Analyzing topics and authors in chat logs for crime investigation[J]. Knowledge & information systems, 2014, 39(2):351 – 381.

[23] CHEN C, REN J. Forum Latent Dirichlet Allocation for user interest discovery[J]. Knowledge-based systems, 2017, 126(C):1 – 7.

[24] LIU S N, LIU C, PENG Z, et al. Mining individual learning topics in course reviews based on author topic model[J]. International journal of distance education technologies, 2017, 15(3): 1 – 14.

[25] YANG T, COMAR P M, XU L. Community detection by popularity based models for authored networked data[C]// IEEE/ACM international conference on advances in social networks analysis and mining. New York: IEEE, 2013:74 – 81.

[26] MORCHID M, PORTILLA Y, JOSSELIN D, et al. An author-topic based approach to cluster tweets and mine their location[J]. Procedia environmental sciences, 2015, 27(7): 26 – 29.

[27] MUKHERJEE S, BASU G, JOSHI S. Joint author sentiment topic model[C]// Computer security applications conference. New York: IEEE, 2014:90 – 98.

[28] 李春山. 面向社会化媒体内容的若干聚类算法研究[D]. 哈尔滨:哈尔滨工业大学, 2014.

[29] 范长俊. 基于信息交互网络的个体角色识别方法研究[D]. 长沙:国防科学技术大学, 2015.

[30] 罗杰斯. 结合用户及地理信息的图像主题建模[D]. 杭州:浙江大学, 2012.

[31] CHENG K, ZHAN Y, QI M. AL-DDCNN: a distributed crossing semantic gap learning for person re-identification[J]. Concurrency & computation practice & experience, 2017, 29(3):1 – 16.

[32] MORCHID M, DUFOUR R, BOUALLEGUE M, et al. Author-topic based representation of call-center conversations[C]// Spoken language technology workshop. New York: IEEE, 2014:218 – 223.

[33] MORCHID M, DUFOUR R, LINARÈS G, et al. Latent topic model based representations for a robust theme identification of highly imperfect automatic transcriptions[J]. Lecture notes in computer science, 2015, 9042(2) :596 – 605.

[34] LUO W, LI H, LIU G, et al. Semantic annotation of satellite images using author-genre-topic model[J]. IEEE transactions on geoscience & remote sensing, 2013, 52(2):1356 – 1368.

[35] ZHU X D, YAO Y, LIU Z J, et al. Activity clustering for online anomaly detection[J]. Journal of computers, 2011, 6(6):441 – 453.

[36] 史庆伟，乔晓东，徐硕,等. 作者主题演化模型及其在研究兴趣演化分析中的应用[J]. 情报学报, 2013, 32(9):912 – 919.

[37] 刘智超,卢美莲. 基于混合模型的学术论文推荐方法[EB/OL]. [2019 – 01 – 18]. http://www.paper.edu.cn/releasepaper/content/201411 – 282.

[38] WANG J, HU X, TU X, et al. Author-conference topic-connection model for academic network search[C]// ACM international conference on information and knowledge management. New York: ACM, 2012:2179 – 2183.

[39] HA J K, AN J Y, YOO K J, et al. Exploring the leading authors and journals in major topics by citation sentences and topic modeling[C]// BIRNDL 2016 joint workshop on bibliometric-enhanced information retrieval and NLP for digital libraries. New York: ACM, 2016: 42 – 50.

[40] 李杰，王小伟. 基于作者主题模型的遥感图像自动类别标注方法[J]. 计算机应用与软件, 2013, 26(10):263 – 265.

[41] LUO W, LI H, LIU G, et al. Global salient information maximization for saliency detection[J]. Signal processing image communication, 2012, 27(3):238 – 248.

[42] LUO W, LI H, LIU G. Automatic annotation of multispectral satellite images using author-topic model[J]. IEEE geoscience & remote sensing letters, 2012, 9(4):634 – 638.

[43] BERTOLERO M A, YEO B T, D'ESPOSITO M. The modular and integrative functional architecture of the human brain[J]. Proceedings of the National Academy of Sciences of the United States of America, 2015, 112(49):798 – 807.

[44] 王永贵，张旭，任俊阳,等. 结合微博关注特性的 UF_AT 模型用户兴趣挖掘研究[J]. 计算机应用研究, 2015, 32(7):1982 – 1985.

[45] 王永贵，张丰田，刘雨诗,等. 微博中结合转发特性的用户兴趣话题挖掘方法[J]. 计算机应用研究, 2017, 34(7):2068 – 2071.

[46] 李敬，印鉴，刘少鹏,等. 基于话题标签的微博主题挖掘[J]. 计算机工程, 2015, 41(4):30 – 35.

[47] 王萍. 网络环境下的领域知识挖掘[D]. 上海:华东师范大学, 2010.

[48] ZHONG Y, FAN Y, TAN W, et al. Web service recommendation with reconstructed profile from mashup descriptions[J]. IEEE transactions on automation science & engineering, 2016, 15(2): 468-478.
[49] TANG J, ZHANG J. Modeling the evolution of associated data[J]. Data & knowledge engineering, 2010, 69(9):965-978.
[50] YANG M, MEI J, XU F, et al. Discovering author interest evolution in topic modeling[J]. 2016, 21(7):801-804.
[51] ZHOU X, SHUNXIANG W U, ZHOU X, et al. The biterm author topic in the sentences model for e-mail analysis[J]. IEICE transactions on information & systems, 2017, 100(8):1852-1859.
[52] MCCALLUM A, WANG X, CORRADA-EMMANUEL A. Topic and role discovery in social networks with experiments on enron and academic email[J]. Journal of artificial intelligence research, 2010, 30(2):249-272.
[53] KANG I S, NA S H, LEE S, et al. On co-authorship for author disambiguation[J]. Information processing & management an international journal, 2009, 45(1):84-97.
[54] ZHANG R, SHEN D, KOU Y, et al. Author name disambiguation for citations on the deep web[J]. Lecture notes in computer science, 2010, 6185(10):198-209.
[55] 郑威杰. 科技文献作者消歧方法研究[D]. 杭州:杭州电子科技大学, 2017.
[56] 陈永恒, 左万利, 林耀进. 作者标签主题模型在科技文献中的应用[J]. 计算机应用, 2015, 35(4):1001-1005.
[57] LI C S, YE Y M, ZHANG X F. TPS: an unsupervised web page segmentation algorithm based on DOM tree structure mining[J]. Information, 2012, 1(15): 387-394.
[58] 桂小庆, 张俊, 张晓民, 等. 时态主题模型方法及应用研究综述[J]. 计算机科学, 2017, 44(2):46-55.
[59] DUFOUR R, MORCHID M, PARCOLLET T. Tracking dialog states using an Author-Topic based representation[C]// Spoken language technology workshop. New York: IEEE, 2017:544-551.
[60] LI D C, OKAMOTO J, LEISCHOW S, et al. An author topic analysis of tobacco regulation investigators[C]// Pacific-Asia conference on knowledge discovery and data mining. Cham: Springer, 2014:616-627.
[61] 亓晓青, 景晓军. 应用于微博的LDA模型改进[EB/OL]. [2019-01-18]. http://www.paper.edu.cn/releasepaper/content/201212-118.
[62] 孙国超, 徐硕, 乔晓东. AToT模型可视化工具开发[J]. 情报工程, 2016, 2(4):20-29.
[63] XU S, SHI Q, QIAO X, et al. A dynamic users' interest discovery model with distributed inference algorithm[J]. International journal of distributed sensor networks, 2014, 2014(1):1-11.
[64] HO T, DO P. Analyzing users' interests with the temporal factor based on topic modeling[C]// Asian conference on intelligent information and database systems. Cham: Springer International Publishing, 2015:105-115.
[65] WANG X, MCCALLUM A. Topics over time: a non-Markov continuous-time model of topical trends[C]// ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM, 2006:424-433.
[66] 杨如意. 基于主题模型的文本语义挖掘[D]. 西安:西安电子科技大学, 2015.
[67] 杨如意, 刘东苏, 李慧. 一种融合外部特征的改进主题模型[J]. 现代图书情报技术, 2016, 32(1):48-54.
[68] 余传明, 左宇恒, 郭亚静, 等. 基于复合主题演化模型的作者研究兴趣动态发现[EB/OL]. [2018-05-26]. http://kns.cnki.net/kcms/detail/37.1389.N.20180419.1330.002.html.
[69] LI C S, CHEUNG W K, YE Y, et al. The Author-Topic-Community model: a generative model relating authors' interests and their community structure[C]// International conference on advanced data mining and applications. Berlin: Springer, 2012:753-765.
[70] LI C, CHEUNG W K, YE Y, et al. The Author-Topic-Community model for author interest profiling and community discovery[J]. Knowledge & information systems, 2015, 44(2):359-383.
[71] YAN E, DING Y, MILOJEVIC S, et al. Topics in dynamic research communities: an exploratory study for the field of information retrieval[J]. Journal of informetrics, 2012, 6(1):140-153.
[72] DING Y. Topic-based PageRank on author cocitation networks[J]. Journal of the American Society for Information Science and Technology. 2011, 62(3): 449-466.
[73] YANG J, ZENG J, CHEUNG W K. Multiplex Topic Models[C]// Pacific-Asia conference on knowledge discovery and data mining. Berlin Heidelberg: Springer, 2013:568-582.
[74] 江雨燕, 李平, 王清, 等. 融合DSTM和USTM方法的主题模型[J]. 计算机科学与探索, 2014, 8(5):630-639.
[75] YANG Z, HONG L, DAVISON B D. Academic network analysis: a joint topic modeling approach[C]//IEEE/ACM international conference on advances in social networks analysis and mining. New York: IEEE, 2014:324-333.
[76] MCCALLUM A, CORRADA-EMMANUEL A, WANG X. Topic and role discovery in social networks[J]. IJCAI, 2005, 30(2): 786-791.
[77] NAVEED N, SIZOV S, STAAB S. ATT: analyzing temporal dynamics of topics and authors in social media[C]// Proceedings of the 3rd international web science conference. New York: ACM, 2011:1-7.
[78] YANG Q Q, LI W J. The LDA Topic Model Extension Study[J]. Logistics engineering, management and computer science, 2015, 169(15): 857-860.
[79] PODDAR L, HSU W, LEE M L. Author-aware aspect topic senti-

ment model to retrieve supporting opinions from reviews[C]// Conference on empirical methods in natural language processing. Berlin Heidelberg:Springer,2017:472-481.

[80] 万路康,章倩雯,谢瑾奎. 社会网络下的基于主题概率的影响从众性模型和分析[J]. 小型微型计算机系统, 2017, 38(2): 277-281.

[81] SCHNEIDER K M. Weighted average pointwise mutual information for feature selection in text categorization[M]// Knowledge discovery in databases: PKDD 2005. Berlin Heidelberg: Springer, 2005:252-263.

[82] WALLACH H M, MURRAY I, SALAKHUTDINOV R, et al. Evaluation methods for topic models[C]// International conference on machine learning. New York:ACM, 2009:1105-1112.

[83] BUNTINE W. Estimating likelihoods for topic models[M]. Advances in machine learning. Berlin Heidelberg: Springer, 2009: 51-64.

[84] XIE X, LI L, ZHANG Z, et al. Back-buy prediction based on TriFG[C]// ACM SIGKDD workshop on mining data semantics. New York: ACM, 2012:1-8.

[85] ZHANG T, LUO W. Image quality assessment using author topic model[C]// International Conference on Information Technology & Applications. New York: IEEE, 2013: 63-65.

[86] XU Z, RU L, XIANG L, et al. Discovering user interest on twitter with a modified author-topic model[C]// IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology. New York: IEEE, 2011:422-429.

[87] 余攀. 基于话题模型的教育领域微博账号萃取[D]. 武汉:华中师范大学, 2017.

[88] 何力,贾焰,韩伟红,等. 基于用户主题模型的微博用户兴趣挖掘[J]. 中国通信, 2014, 11(8): 131-144.

[89] 裯良. 基于主题模型的企业微博推荐方法研究与实现[D]. 合肥:安徽大学, 2016.

[90] TU Y, JOHRI N, DAN R, et al. Citation author topic model in expert search[C]//International conference on computational linguistics: Posters. New York: ACM, 2010:1265-1273.

[91] KATARIA S, MITRA P, CARAGEA C, et al. Context sensitive topic models for author influence in document networks[C]// International joint conference on artificial intelligence. California: AAAI, 2011:2274-2280.

[92] YAN L, NICULESCU-MIZIL A, GRYC W. Topic-link LDA: joint models of topic and author community[C]// International conference on machine learning. New York: ACM, 2009:665-672.

作者贡献说明:

徐涵:确定选题方向,数据收集、处理与分析,论文撰写与修改;

刘小平:提出修改意见。

A Review of Methods and Applications for Author-Topic Model and Its Improved Models

Xu Han[1,2] Liu Xiaoping[1,2]

[1] National Science Library, Chinese Academy of Sciences, Beijing 100190

[2] Department of Library, Information and Archives Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190

Abstract: [**Purpose/significance**] Author-Topic model, as a new probabilistic model which has a high degree of attention in computer science, has been widely applied in text mining, natural language processing and other fields in recent years. This paper analyzes the ideas and applications of AT model and its improved models to grasp its research status and provide reference and ideas for researchers in computer science, library and information science or some other related fields. [**Method/process**] Using data sets on Web of Science Core Collection, DBLP and CNKI (China Academic Journals Full-text Database), a literature collection on Author-Topic model and its improved models is constructed through the establishment of retrieval rules, data de-duplication, artificial judgment and other operations. This paper summarizes the existing research based on literature analysis method from the perspective of the application process of the model. [**Result/conclusion**] The results show that the existing related research has formed a comparatively complete analysis process and the improvement angle and application area of the models are increasingly diversified. However, some problems, such as performance optimization, standardization and perfection and further application in the field of library and information science, still need to be explored in depth.

Keywords: Author-Topic model topic evolution community detection model evolution